RICH FEATURES FOR PERCEPTUAL QUALITY ASSESSMENT OF UGC VIDEOS

SUPPLEMENTARY MATERIALS

Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck,
Balu Adsumilli, Peyman Milanfar, Feng Yang
Google Inc.

1 Subjective test platform for UGC-VQ dataset

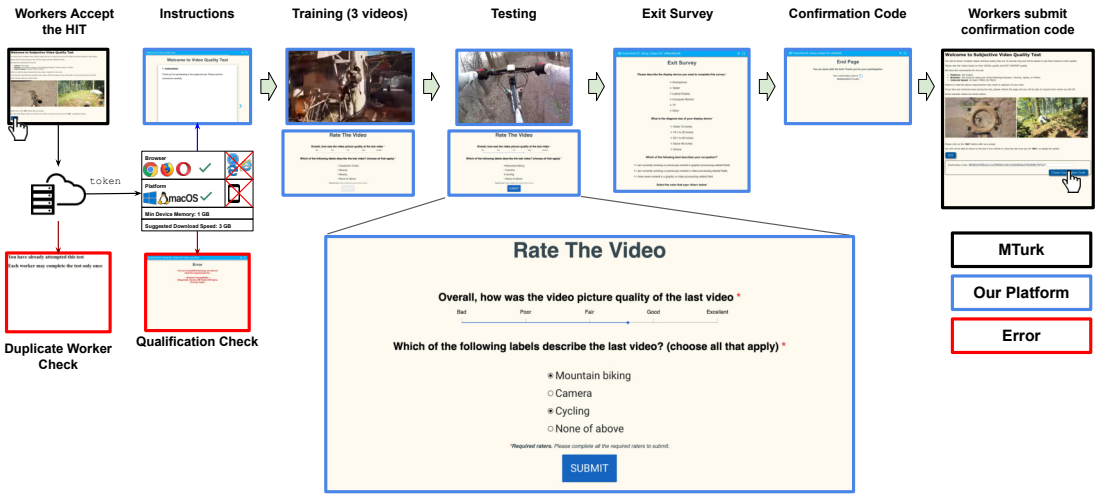


Figure 1: Subjective test platform overview

We utilized Mechanical Turk (MTurk) to delegate the recruitment and payment process. The external workers were redirected to our testing platform, which is a standalone web application. Figure 1 shows the step by step flow of the MTurk workers' journey through our subjective testing and the below steps highlight the details of the critical parts of the testing:

1. When MTurk workers entered our platform, we ran basic qualification checks, which was used to reject unqualified workers (e.g., poor Internet bandwidth or the display resolution is smaller than 700P), and also displayed detailed instructions about the test.
2. While Step 1 was in progress, we downloaded the videos to the workers' local drive to decouple network related issues from the quality assessment.
3. After watching the entire clip, workers were asked to rate the clip on the picture quality and to choose all relevant labels. Each clip had a different set of labels as options for the second question.
4. At the end of the test, exit survey questions were asked to collect basic demographic information about the workers.

Since our corpus was larger than what a single worker could watch within a reasonable time frame, we sampled subset of videos from the testing corpus. We showed three training videos in the beginning to get the workers familiar with our platform and to calibrate workers with the same quality range because each subset had different ranges of quality. The actual stimuli were randomly shuffled while making sure the variants from the same video ID were not shown consecutively.

2 Comparison of aggregation models

We compared the three feature aggregation models (AvgPool, LSTM, and ConvLSTM) by their MOS correlations with all combinations of features. As shown in Table 1, for all three models the combined features performs better than a single feature, and CP+CT+DT has the highest correlation on all three models. It means the three extracted features are all highly related to perceptual quality. AvgPool has better correlations than LSTM and ConvLSTM in most cases, which suggests that the majority of UGC videos still have relatively consistent quality. ConvLSTM outperforms LSTM on single features, but the advantage decreases in combined feature cases. For fine-tuning on such small datasets (around 1k samples, which is common for video quality assessment), complicated models may not perform better than relatively simple models. Thus in this paper, we choose AvgPool as our default AggregationNet.

Feature	AvgPool			LSTM			ConvLSTM		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
CP	0.770	0.785	0.408	0.674	0.684	0.473	0.717	0.727	0.445
CT	0.628	0.628	0.495	0.584	0.574	0.518	0.621	0.618	0.502
DT	0.726	0.744	0.434	0.722	0.727	0.443	0.752	0.758	0.422
CP+CT	0.787	0.801	0.395	0.739	0.741	0.430	0.734	0.732	0.436
CP+DT	0.790	0.802	0.391	0.768	0.773	0.408	0.774	0.778	0.407
CT+DT	0.750	0.767	0.421	0.730	0.739	0.437	0.730	0.735	0.437
CP+CT+DT	**0.802**	**0.816**	**0.382**	0.767	0.771	0.411	0.760	0.764	0.418

Table 1: Comparison of different temporal aggregation models on YT-UGC original MOS.

3 Comparison on model complexity

Our default backbone model for ContentNet and DistortionNet is EfficientNet-b0 (pretrained on ImageNet). In this section, we will investigate whether a more complex model could achieve better performance. We use EfficientNet-b7 (pretrained on ImageNet) as the alternative model, to evaluate the impact of model complexity on overall quality prediction. The complexity of EfficientNet-b0 and EfficientNet-b7 are 0.39B FLOPs and 37B FLOPs, respectively.

First, we use Efficient-b0 and Efficient-b7 as the backbone for ContentNet to extract content features. These features are then combined with other features in our AggregationNet (AvgPool) to predict the YT-UGC MOS (Table 2) and the UGC-VQ DMOS (Table 3). We can see EfficientNet-b7 does not achieve correlations higher than EfficientNet-b0 (actually slightly worse) in single feature model (CT) as well as the combined feature models. This means that more complicated models do not always provide more useful features. Thus, EfficientNet-b0 is the best option for our CoINVQ model.

Feature	EfficientNet-b0			EfficientNet-b7		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
CT	0.628	0.628	0.495	0.615	0.607	0.498
CP+CT	0.787	0.801	0.395	0.774	0.778	0.405
CT+DT	0.750	0.767	0.421	0.752	0.770	0.421
CP+CT+DT	**0.802**	**0.816**	**0.382**	0.796	0.806	0.387

Table 2: Comparison of different ContentNet backbone on YT-UGC original MOS.

Feature	EfficientNet-b0			EfficientNet-b7		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
CT	0.584	0.526	0.106	0.357	0.373	0.130
CP+CT	**0.672**	**0.613**	**0.192**	0.652	0.602	0.214
CT+DT	0.390	0.387	0.201	0.334	0.347	0.215
CP+CT+DT	0.539	0.522	0.203	0.497	0.481	0.223

Table 3: Comparison of different ContentNet backbone on UGC-VQ DMOS.

Model	PLCC	SROCC	RMSE
EfficientNet-b0(frozen)	0.624	0.612	0.509
EfficientNet-b0(finetune)	0.671	0.690	0.474
EfficientNet-b7(frozen)	0.5447	0.5435	0.5364
EfficientNet-b7(finetune)	0.6035	0.5961	0.5147
CoINVQ (CT)	0.628	0.628	0.495
CoINVQ (CP+CT)	0.787	0.801	0.395
CoINVQ (CT+DT)	0.750	0.767	0.421
CoINVQ (CP+CT+DT)	0.802	0.816	0.382

Table 4: Comparison of EfficientNet-b0 and EfficientNet-b7 on YT-UGC original MOS.

We also evaluate the direct influence of the model size on the overall quality prediction performance. We test two frame-based models: EfficientNet-b0 and EfficientNet-b7 (pretrained on ImageNet). Each model has two retraining strategies: frozen (model weights with a trainable head) and funetuned (all model weights are trainable). Table 4 shows the correlations on YT-UGC original MOS, where we still use the 5-fold cross-validation with consistent splits for all tests and report average results over the test folds. We can see again EfficientNet-b7 did not give better correlation, and is much worse than our CoINVQ models. This suggests that more comprehensive features can bring more gains than more complicated models for video quality assessment.

To evaluate the benefit of retraining on YT8M content labels, we ran another experiment to compare the performance of the ContentNet features (retrained on YT8M), and features from EfficientNet (pretrained on ImageNet). These content features are directly used to train (not combined with other features) on YT-UGC original MOS. Then, the trained models are used to predict DMOS for UGC-VQ dataset without retraining. Both ContentNet and EfficientNet have two backbone networks (b0 and b7). Results are shown in Table 5. We found out that after retraining on the YT8M content labels, the models achieve significantly better correlations than the models just trained on ImageNet. This means that retraining on UGC content labels is necessary for the UGC quality assessment.

Dataset	ContentNet-b0			ContentNet-b7			EfficientNet-b0			EfficientNet-b7		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
YT-UGC MOS	0.628	0.628	0.495	0.615	0.607	0.498	0.586	0.577	0.517	0.583	0.565	0.519
UGC-VQ DMOS	0.584	0.526	0.106	0.357	0.373	0.130	0.485	0.395	0.119	0.392	0.396	0.122

Table 5: Performance of different content features used in CoINVQ(CT) model on YT-UGC MOS and UGC-VQ DMOS.

4 Scatter Plots for predictions and ground truth

4.1 YT-UGC original MOS: Table 6

4.2 KoNViD-1k original MOS (not retrained): Table 7

4.3 UGC-VQ DMOS (not retrained): Table 8

5 More CoINVQ diagnosis reports on YT-UGC dataset

Here we list CoINVQ diagnosis reports for samples form YT0-UGC dataset (sorted by MOS in ascending order). Compression level (in [0, 1]), top-10 content labels, and top 5 distortion types are reported (sorted by predicted probability in descending order). Quality scores predicted by a single feature (CP, CT, DT) and combined features (CP+CT+DT) are shown as well as ground truth MOS. We can see for most samples, predicted overall quality scores are close to MOS, and quality indicators are helpful to understand the generic video quality. Several videos (e.g., Sports_720P-069c) whose scores predicted by single features (e.g., CT) are closer to the ground truth than the scores predicted by combined features, which suggests that there are still room of improvement for feature aggregation.

All scores are averaged for the entire 20s videos, so selected frames may not fully represent the overall video quality. Full videos can be accessed by clicking corresponding vids. Due to the license issue, the link will be disabled if the video was removed from YT-UGC dataset.

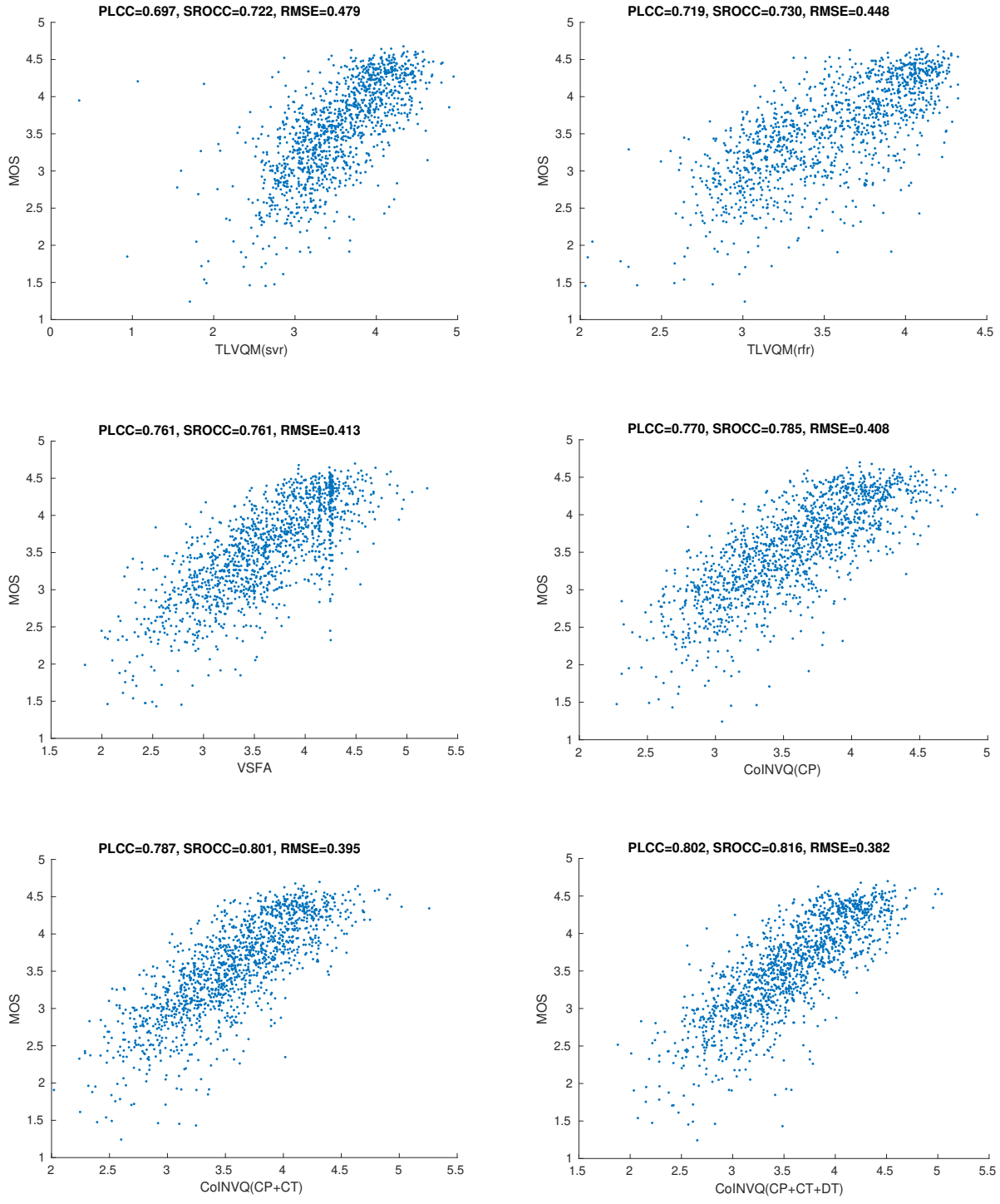


Table 6: Scatter plots for metrics on YT-UGC MOS.

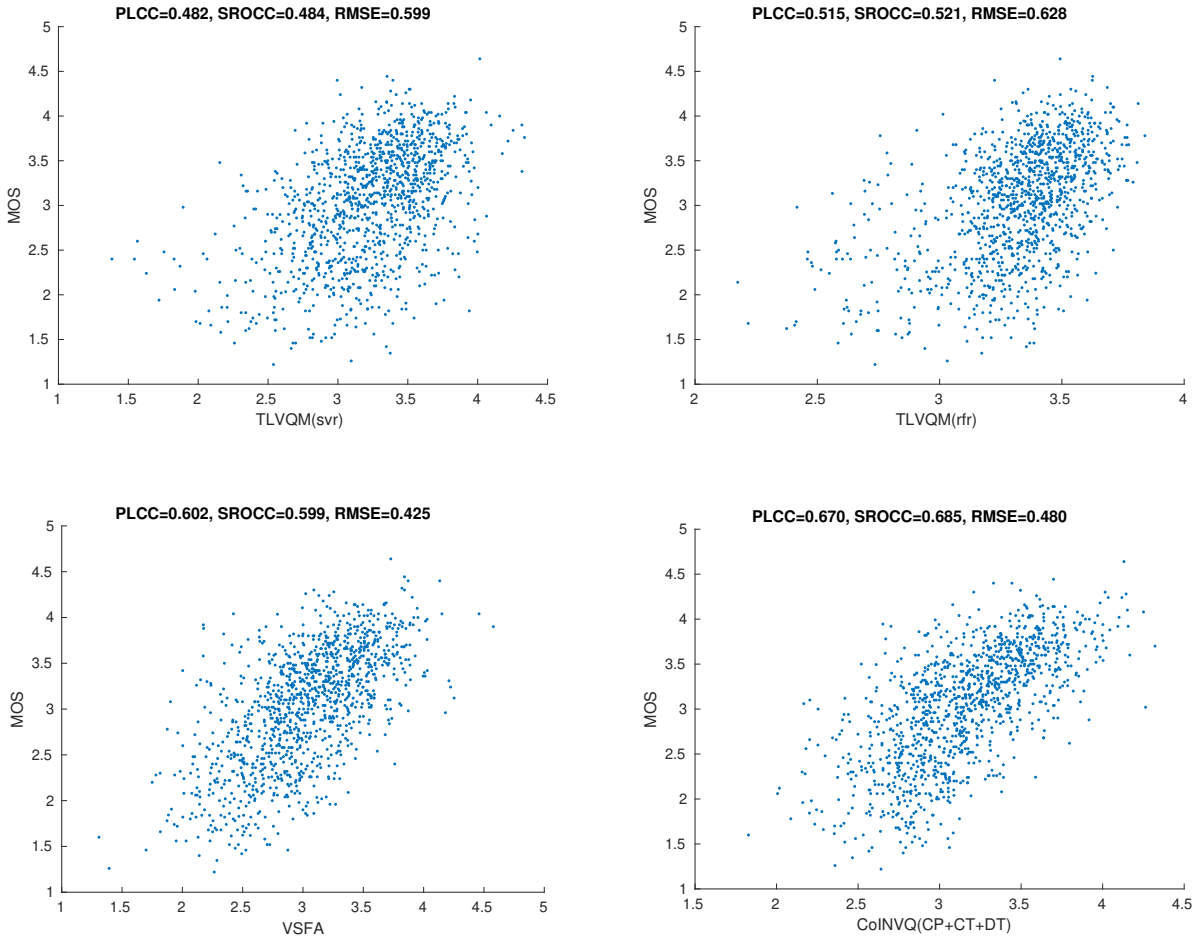


Table 7: Scatter plots for metrics on KoNViD-1k MOS (not retrained).

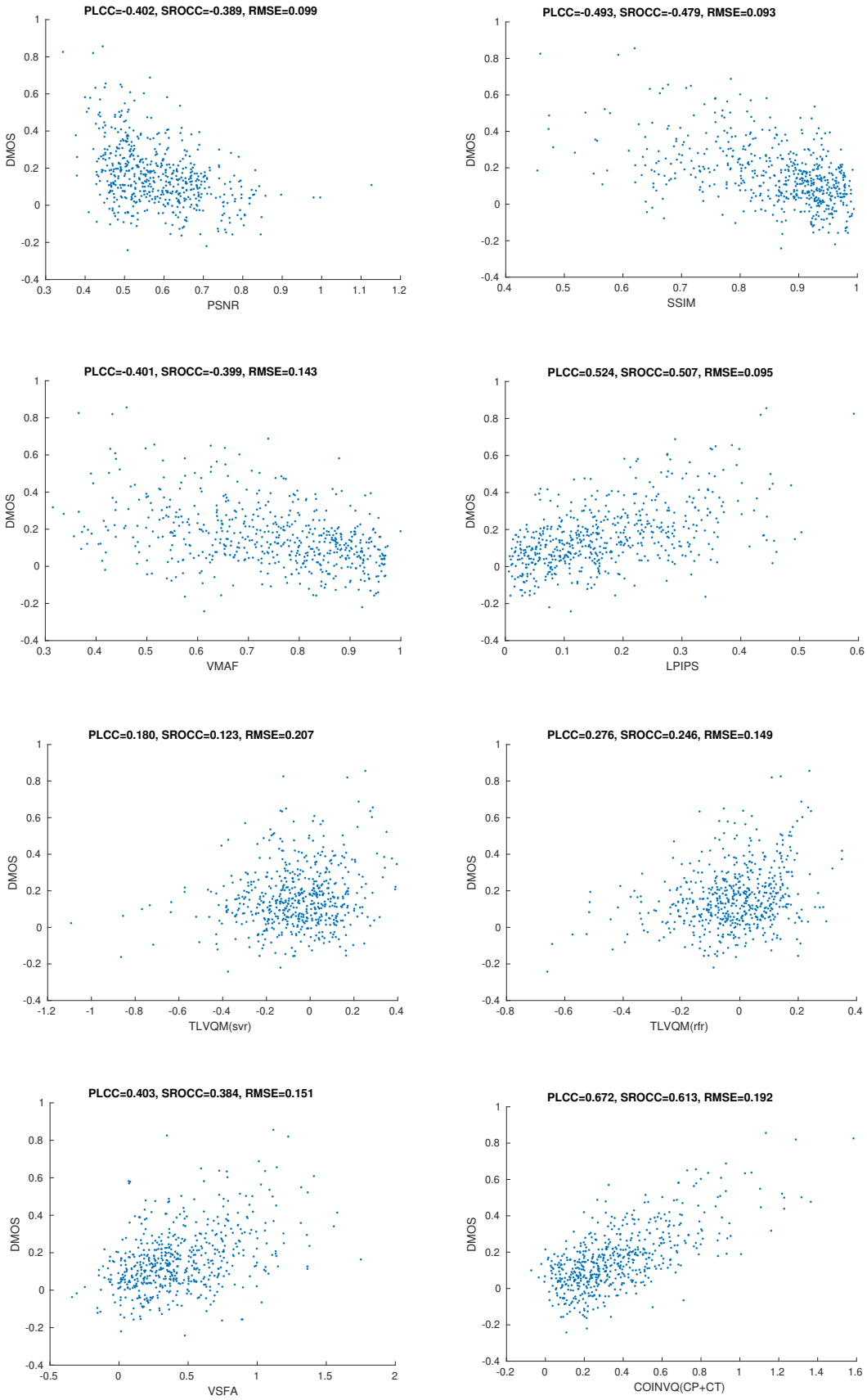
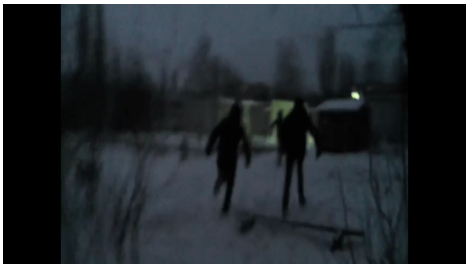


Table 8: Scatter plots for metrics on UGC-VQ DMOS (not retrained).

Video link: HDR_1080P-1e5b

CoINVQ diagnosis report	
Compression level	0.067
Content labels	Video game, Food, Vehicle, Music video, Minecraft, Animal, Trailer (promotion), Car, Guitar, Pet
Distortion types	Lens blur, Multiplicative noise, Gaussian blur, Denoise, JPEG2000
(CP, CT, DT)	(3.048, 2.619, 2.792)
CP+CT+DT	2.654
MOS (ground truth)	1.242

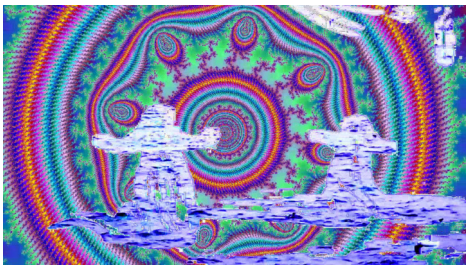


Video link: TelevisionClip_1080P-3e42

CoINVQ diagnosis report	
Compression level	0.225
Content labels	Video game, Trailer (promotion), Music video, Vehicle, Animal, Car, Concert, Musician, Performance art, Pet
Distortion types	Denoise, Lens blur, Jitter, Color quantization, Color saturation 2
(CP, CT, DT)	(2.861, 2.599, 2.713)
CP+CT+DT	2.621
MOS (ground truth)	1.989

Video link: LiveMusic_360P-5b57

CoINVQ diagnosis report	
Compression level	0.794
Content labels	Vehicle, Musician, Guitar, String instrument, Car, Food, Video game, Motorcycle, Animal, Music video
Distortion types	Color saturation 2, Multiplicative noise, Gaussian blur, Lens blur, Denoise
(CP, CT, DT)	(2.605, 3.301, 2.826)
CP+CT+DT	2.718
MOS (ground truth)	2.773



Video link: CoverSong_720P-10f1

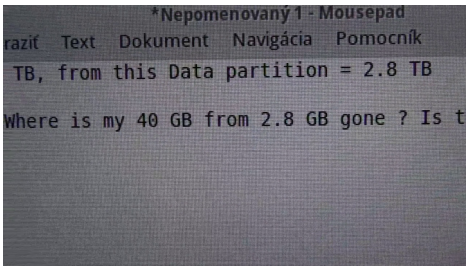
CoINVQ diagnosis report	
Compression level	0.155
Content labels	Cartoon, Toy, Dance, Performance art, Animation, Animal, Slam dunk, Video game, Doll, Art
Distortion types	Jitter, Contrast change, Color block, Darken, White noise in color component
(CP, CT, DT)	(3.383, 3.835, 2.412)
CP+CT+DT	2.248
MOS (ground truth)	2.805

Video link: Vlog_1080P-45c9

CoINVQ diagnosis report	
Compression level	0.069
Content labels	Vehicle, Car, Video game, Animal, Pet, Train, Rail transport, Fishing, Food, Musical keyboard
Distortion types	Quantization, Denoise, Lens blur, JPEG2000, Gaussian blur
(CP, CT, DT)	(3.570, 3.268, 3.164)
CP+CT+DT	3.296
MOS (ground truth)	2.888

Video link: Gaming_360P-586d

CoINVQ diagnosis report	
Compression level	0.145
Content labels	Video game, Guitar, Musician, String instrument, Strategy video game, Warcraft, World of Warcraft, Minecraft, Trailer (promotion), Call of Duty
Distortion types	Gaussian blur, Pixelate, Color saturation 2, Mean shift, Color saturation 1
(CP, CT, DT)	(3.465, 3.509, 3.173)
CP+CT+DT	3.228
MOS (ground truth)	3.026



Video link: HowTo_720P-3813

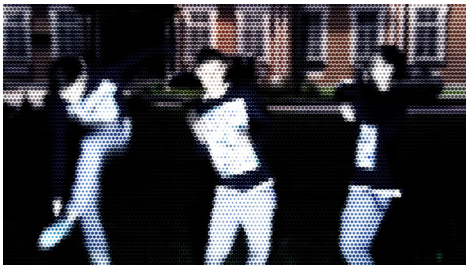
CoINVQ diagnosis report	
Compression level	0.137
Content labels	Vehicle, Video game, Car, Musician, Cartoon, Guitar, Dance, Food, Mobile phone, Photography
Distortion types	Denoise, Lens blur, Quantization, JPEG2000, Color shift
(CP, CT, DT)	(3.228, 3.770, 3.870)
CP+CT+DT	3.738
MOS (ground truth)	3.072

Video link: Sports_480P-0623

CoINVQ diagnosis report	
Compression level	0.197
Content labels	Video game, Vehicle, Animal, Trailer (promotion), Association football, Guitar, Pet, Car, Food, Dance
Distortion types	White noise in color component, Unknown, Denoise, Contrast change, Lens blur
(CP, CT, DT)	(2.924, 3.241, 2.958)
CP+CT+DT	2.943
MOS (ground truth)	3.106

Video link: LiveMusic_480P-1a91

CoINVQ diagnosis report	
Compression level	0.147
Content labels	Video game, Concert, Vehicle, Musician, Performance art, Car, Trailer (promotion), Guitar, Dance, Musical ensemble
Distortion types	Gaussian blur, Unknown, Pixelate, Lens blur, Denoise
(CP, CT, DT)	(3.294, 3.092, 2.960)
CP+CT+DT	3.086
MOS (ground truth)	3.173



Video link: Vlog_2160P-217c

CoINVQ diagnosis report	
Compression level	0.021
Content labels	Concert, Performance art, Musician, Dance, Musical ensemble, Vehicle, Car, Video game, Guitar, Cartoon
Distortion types	Gaussian blur, Denoise, Quantization, Non-eccentricity patch, Multiplicative noise
(CP, CT, DT)	(3.526, 3.190, 3.476)
CP+CT+DT	3.494
MOS (ground truth)	3.202

Video link: Animation_480P-6ff4

CoINVQ diagnosis report	
Compression level	0.543
Content labels	Video game, Minecraft, Strategy video game, Food, World of Warcraft, Nature, Combat, Animal, Dance, Association football
Distortion types	Jitter, Non-eccentricity patch, Unknown, Lens blur, Impulse noise
(CP, CT, DT)	(3.361, 3.675, 3.245)
CP+CT+DT	3.379
MOS (ground truth)	3.290

Video link: Sports_480P-6508

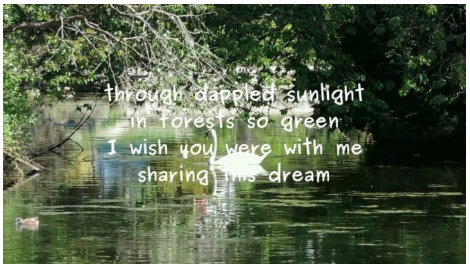
CoINVQ diagnosis report	
Compression level	0.042
Content labels	Vehicle, Video game, Car, Mobile phone, Animal, Smartphone, Gadget, Motorsport, Pet Food
Distortion types	Jitter, Color quantization, Lens blur, Denoise, Pixelate
(CP, CT, DT)	(2.932, 2.934, 3.316)
CP+CT+DT	3.476
MOS (ground truth)	3.448

Video link: NewsClip_720P-4603

CoINVQ diagnosis report	
Compression level	0.190
Content labels	Vehicle, Car, Video game, Train, Food, Rail transport, Musical keyboard, Animal, Railroad car, Pet
Distortion types	Lens blur, Quantization, JPEG2000, Denoise, Color shift
(CP, CT, DT)	(3.483, 3.484, 3.562)
CP+CT+DT	3.436
MOS (ground truth)	3.454

Video link: Vlog_720P-329f

CoINVQ diagnosis report	
Compression level	0.283
Content labels	Animal, Pet, Guitar, Vehicle, Food, Video game, Train, Association football, String instrument, Orchestra
Distortion types	Lens blur, Quantization, JPEG2000, Denoise, Mean shift
(CP, CT, DT)	(3.654, 3.403, 3.828)
CP+CT+DT	3.747
MOS (ground truth)	3.700



Video link: LyricVideo_720P-0ae4

CoINVQ diagnosis report	
Compression level	0.102
Content labels	Vehicle, Animal, Car, Pet, Slam dunk, Food, Orchestra, Musical keyboard, Fishing, Outdoor recreation
Distortion types	Lens blur, Denoise, JPEG2000, Quantization, Mean shift
(CP, CT, DT)	(3.778, 3.348, 4.032)
CP+CT+DT	3.824
MOS (ground truth)	3.717

Video link: Vlog_2160P-4655

CoINVQ diagnosis report	
Compression level	0.016
Content labels	Video game, Strategy video game, Fishing, Pet, World of Warcraft, The Sims, Animal, Fish, Vehicle, Trailer (promotion)
Distortion types	Gaussian blur, Multiplicative noise, Color saturation 2, Lens blur, Denoise
(CP, CT, DT)	(3.666, 3.545, 3.639)
CP+CT+DT	3.728
MOS (ground truth)	3.887

Video link: Sports_720P-0b9e

CoINVQ diagnosis report	
Compression level	0.226
Content labels	Animal, Vehicle, Pet, Musical keyboard, Rail transport, Car, Food, Train, Railroad car, Orchestra
Distortion types	Impulse noise, Mean shift, Unknown, Color saturation 1, Jitter
(CP, CT, DT)	(3.723, 3.961, 3.556)
CP+CT+DT	3.646
MOS (ground truth)	3.942

Video link: LyricVideo_1080P-5461

CoINVQ diagnosis report	
Compression level	0.093
Content labels	Video game, Music video, Trailer (promotion), Concert, Car, Vehicle, Performance art, Musician, Guitar, Animal
Distortion types	Denoise, Gaussian blur, Quantization, Mean shift, Lens blur
(CP, CT, DT)	(3.516, 3.231, 3.304)
CP+CT+DT	3.367
MOS (ground truth)	4.031

Video link: LiveMusic_1080P-6b1c

CoINVQ diagnosis report	
Compression level	0.124
Content labels	Concert, Musician, Video game, Vehicle, Guitar, Musical ensemble, Call of Duty, Drum kit, Performance art, Car
Distortion types	Contrast change, Non-eccentricity patch, Impulse noise, Color quantization, White noise in color component
(CP, CT, DT)	(4.164, 3.606, 4.245)
CP+CT+DT	4.100
MOS (ground truth)	4.255

Video link: MusicVideo_480P-5461

CoINVQ diagnosis report	
Compression level	0.066
Content labels	Vehicle, Music video, Cartoon, Car, Video game, Art, Drawing, Mobile phone, Food, Smartphone
Distortion types	Lens blur, Denoise, Quantization, JPEG2000, Color saturation 2
(CP, CT, DT)	(3.802, 3.751, 3.817)
CP+CT+DT	3.921
MOS (ground truth)	4.272

Video link: Sports_2160P-349c

CoINVQ diagnosis report	
Compression level	0.020
Content labels	Vehicle, Mobile phone, Car, Gadget, Airplane, Smartphone, Aircraft, Aircraft, Animal, Pokemon (video game series), Pet
Distortion types	Quantization, Mean shift, JPEG2000, Denoise, Lens blur
(CP, CT, DT)	(4.237, 4.079, 4.936)
CP+CT+DT	4.526
MOS (ground truth)	4.302

Video link: Sports_1080P-43e2

CoINVQ diagnosis report	
Compression level	0.075
Content labels	Association football, Vehicle, Food, Animal, Video game, Dance, Pet, Train, Minecraft, Car
Distortion types	Lens blur, Impulse noise, Denoise, Color saturation 2, Color shift
(CP, CT, DT)	(4.396, 4.300, 4.092)
CP+CT+DT	4.254
MOS (ground truth)	4.414

Video link: Sports_720P-3338

CoINVQ diagnosis report	
Compression level	0.096
Content labels	Dance, Association football, Cycling, Video game, Choir, Musical ensemble, Animal, Vehicle, Basketball moves, Sports game
Distortion types	Color shift, Contrast change, White noise in color component, Denoise, Mean shift
(CP, CT, DT)	(4.208, 3.585, 4.293)
CP+CT+DT	4.378
MOS (ground truth)	4.483

Video link: HDR_1080P-2d32

CoINVQ diagnosis report	
Compression level	0.014
Content labels	Food, Train, Minecraft, Vehicle, Rail transport, Video game, Musical keyboard, Railroad car, Engine, Car
Distortion types	Lens blur, Mean shift, Quantization, Denoise, JPEG2000
(CP, CT, DT)	(4.162, 4.135, 4.596)
CP+CT+DT	4.286
MOS (ground truth)	4.503

Video link: Gaming_1080P-6dc6

CoINVQ diagnosis report	
Compression level	0.005
Content labels	Video game, Strategy video game, Minecraft, Vehicle, Combat, Cartoon, Warcraft, Battlefield , Call of Duty, World of Warcraft
Distortion types	White noise in color component, Color shift, Contrast change, Denoise, Mean shift
(CP, CT, DT)	(4.312, 4.754, 4.621)
CP+CT+DT	4.778
MOS (ground truth)	4.602